

# Optimal Parameter Settings for the $(1 + (\lambda, \lambda))$ Genetic Algorithm\*

Benjamin Doerr, École Polytechnique, France

August 1, 2016

## Abstract

The  $(1 + (\lambda, \lambda))$  genetic algorithm is one of the few algorithms for which a super-constant speed-up through the use of crossover could be proven. So far, this algorithm has been used with parameters based also on intuitive considerations. In this work, we rigorously regard the whole parameter space and show that the asymptotic time complexity proven by Doerr and Doerr (GECCO 2015) for the intuitive choice is best possible among all settings for population size, mutation probability, and crossover bias.

## 1 Introduction

The  $(1 + (\lambda, \lambda))$  genetic algorithm ( $(1 + (\lambda, \lambda))$  GA) was first proposed in [DDE13] (see [DDE15] for the journal version). It is a simple evolutionary algorithm that uses a biased crossover with a parent individual in a way that can be interpreted as a repair mechanism. It was the first (unbiased in the sense of Lehre and Witt [LW12]) evolutionary algorithm to provably optimize any ONEMAX test function in time asymptotically smaller than the famous  $\Theta(n \log n)$  barrier [DD15b], but showed a favorable performance in experiments also for several other classic test functions [DDE15] and combinatorial optimization problems [GP14, MB15]. This algorithm (together with, e.g., [JW02, FW04, Sud05, DHK12]) also is one of the still surprisingly few examples where crossover could be rigorously proven to useful.

One difficulty when using the  $(1 + (\lambda, \lambda))$  GA is that it comes with several parameters, namely an offspring population size  $\lambda$ , a mutation probability  $p$ , and a crossover bias  $c$ . In all previous works, these parameters were chosen by combining rigorous and intuitive arguments (see Section 2). While the results, e.g., an  $O(n\sqrt{\log n})$  runtime for all ONEMAX functions in the first paper [DDE13], indicate that these intuitive choices were not too bad, all existing work leaves open the possibility that completely different parameter choices give an even better performance.

For this reason, in this work we rigorously prove a lower bound valid for the whole 3-dimensional parameter space. Our lower bound coincides with the runtime proven in [DD15b] for the intuitive choices taken there. Consequently, these parameter choices

---

\*A short version of this paper with many proofs omitted appeared at GECCO'16.

were optimal. As a side product of this result, we also see that not many other parameter choices can lead to this optimal runtime. We have to defer the precise statement of our results (Theorem 7) to a point where the algorithms and its parameters have been made precise.

From a broader perspective, our results and in particular the partial results that lead to it, give a clearer picture on how to choose the parameters in the  $(1 + (\lambda, \lambda))$  GA, also for optimization problems beyond the ONEMAX test function class (see the Conclusion section).

From the methodological standpoint, this is one of very few theoretical works that analyze evolutionary algorithms involving more than one parameter. We observe that the parameters do not have an independent influence on the runtime, but that they interact in a difficult to foresee manner. A similar observation was made in [GW15], who proved for the  $(1 + \lambda)$  EA that the mutation probability has a decisive influence on the performance when the population size  $\lambda$  is asymptotically smaller than the cut-off point  $\ln(n) \ln \ln(n) / \ln \ln \ln(n)$ , where as it have almost no influence when  $\lambda = \omega(\ln(n) \ln \ln(n) / \ln \ln \ln(n))$ . Such non-separable parameter influences, naturally, makes the analysis of a multi-dimensional parameter space more difficult. A second difficulty we had to overcome is that, while only few parameter configuration yields the asymptotically optimal runtime, a quite large set of combinations including some that are far from the optimal ones still lead to a runtime very close to the optimal one (see the remark at the end of Section 5.1). While this is good from the application point of view (missing the absolutely optimal parameters is less harmful), from the viewpoint of proving our results it means that there is not much room for non-sharp estimates. Overcoming these difficulties, we are also optimistic that this work helps future work in the analysis of multi-dimensional parameter spaces.

## 2 The $(1 + (\lambda, \lambda))$ GA

The  $(1 + (\lambda, \lambda))$  GA is a fairly simple evolutionary algorithm using crossover. It was introduced in [DDE13, DDE15], some experimental results can be found in [GP14]. Its pseudo-code is given in Algorithm 1.

The  $(1 + (\lambda, \lambda))$  GA is initialized with a solution candidate drawn uniformly at random from  $\{0, 1\}^n$ . It then proceeds in iterations consisting of a mutation, a crossover, and a selection phase. In an important contrast to many other genetic algorithms, the mutation phase *precedes* the crossover phase. This allows to use crossover as a repair mechanism, as we shall discuss in more detail below.

In the *mutation phase* of the  $(1 + (\lambda, \lambda))$  GA, we create  $\lambda$  offspring from the current-best solution  $x$  by applying to it the mutation operator  $\text{mut}_\ell(\cdot)$ , which flip  $\ell$  positions uniformly at random. In other words,  $\text{mut}_\ell(x)$  is a bit-string in which for  $\ell$  random positions  $i$  the entry  $x_i \in \{0, 1\}$  is replaced by  $1 - x_i$ . The *step size*  $\ell$  is chosen randomly according to a binomial distribution  $\mathcal{B}(n, p)$  with  $n$  trials and success probability  $p$ . To ensure that all mutants have the same distance from the parent  $x$ , and thus to not bias the selection by different distances from the parent, the same  $\ell$  is used for all  $\lambda$  offspring. The fitness of the  $\lambda$  offspring is computed and the best one of them,  $x'$ , is selected to take part in the crossover phase. If there are several offspring having maximal fitness, we pick

---

**Algorithm 1:** The  $(1 + (\lambda, \lambda))$  GA, maximizing a given function  $f : \{0, 1\}^n \rightarrow \mathbb{R}$ , with offspring population size  $\lambda$ , mutation probability  $p$ , and crossover bias  $c$ . The mutation operator  $\text{mut}_\ell$  generates an offspring from one parent by flipping exactly  $\ell$  random bits (without replacement). The crossover operator  $\text{cross}_c$  performs a biased uniform crossover, taking bits independently with probability  $c$  from the second argument and with probability  $1 - c$  from the first parent.

---

```

1 Initialization: Choose  $x \in \{0, 1\}^n$  uniformly at random and evaluate  $f(x)$ ;
2 Optimization: for  $t = 1, 2, 3, \dots$  do
3   Mutation phase: Sample  $\ell$  from  $\mathcal{B}(n, p)$ ;
4   for  $i = 1, \dots, \lambda$  do
5      $x^{(i)} \leftarrow \text{mut}_\ell(x)$  and evaluate  $f(x^{(i)})$ ;
6   Choose  $x' \in \{x^{(1)}, \dots, x^{(\lambda)}\}$  with  $f(x') = \max\{f(x^{(1)}), \dots, f(x^{(\lambda)})\}$  u.a.r.;
7   Crossover phase: for  $i = 1, \dots, \lambda$  do
8      $y^{(i)} \leftarrow \text{cross}_c(x, x')$  and evaluate  $f(y^{(i)})$ ;
9   Choose  $y \in \{y^{(1)}, \dots, y^{(\lambda)}\}$  with  $f(y) = \max\{f(y^{(1)}), \dots, f(y^{(\lambda)})\}$  u.a.r.;
10  Selection step: if  $f(y) \geq f(x)$  then  $x \leftarrow y$ ;

```

---

one of them uniformly at random (u.a.r.).

When  $x$  is already close to an optimal solution, the offspring created in the mutation phase are typically all of much worse fitness than  $x$ . Our hope is though that they have discovered some parts of the optimum solution that is not yet reflected in  $x$ . In order to preserve these parts while at the same time not destroying the good parts of  $x$ , the  $(1 + (\lambda, \lambda))$  GA creates in the *crossover phase*  $\lambda$  offspring from  $x$  and  $x'$ . Each one of these offspring is sampled from a uniform crossover with bias  $c$  to take an entry from  $x'$ , that is, each offspring  $y^{(i)} := \text{cross}_c(x, x')$  is created by independently for each position  $j$  setting  $y_j^{(i)} := x'_j$  with probability  $c$  and taking  $y_j^{(i)} := x_j$  otherwise. Again we evaluate the fitness of the  $\lambda$  crossover offspring and select the best one of them, which we denote by  $y$ . If there are several offspring of maximal fitness, we simply take one of them uniformly at random.<sup>1</sup>

Finally, in the *selection step* the previous-best solution  $x$  is replaced by new  $y$  if and only if the fitness of  $y$  is at least as good as the one of  $x$ .

As common in the runtime analysis community, we do not specify a termination criterion. The simple reason is that we study as a theoretical performance measure the expected number of function evaluations that the  $(1 + (\lambda, \lambda))$  GA performs until it evaluates for the first time a search point of maximal fitness (the so-called optimization time). Of course, for an application to a real problem a termination criterion has to be specified.

---

<sup>1</sup>In [DDE13, Section 4.4] and [DDE15] a slightly different selection rule is suggested for the crossover phase, which is more suitable for functions with large plateaus of the same fitness value. Since we consider in this work only the ONEMAX function, for which both algorithms are identical by symmetry reasons, we refrain from stating in Algorithm 1 the slightly more complicated version proposed there, which selects the parent solution  $x$  only if there is no offspring  $\neq x$  of fitness value at least as good as the one of  $x$ .

## Parameter Choices

The  $(1 + (\lambda, \lambda))$  GA comes with a set of parameters, namely the mutation probability  $p$ , the crossover bias  $c$ , and the off-spring population size  $\lambda$ . If  $\ell \sim \mathcal{B}(n, p)$ , then observations that  $\text{cross}_c(x, \text{mut}_\ell(x))$  has the distribution of an individual created from  $x$  via standard bit mutation with mutation rate  $pc$ . Since  $1/n$  is an often preferred choice for the mutation rate, the authors of [DDE13] suggest to choose  $p$  and  $c$  in a way that  $pc = 1/n$ . Note that due to the two intermediate selection steps, the final offspring  $y$  has a very different distribution than standard bit mutation with rate  $pc$  – otherwise the  $(1 + (\lambda, \lambda))$  GA could not obtain runtimes better than  $\Theta(n \log n)$ .

Parameterizing  $p = k/n$ , that is,  $k$  denotes the average number of bits flipped by an application of the mutation operator, the above suggestion is to take  $c = 1/k$ . For these settings, a runtime analysis for the ONEMAX test function in [DDE13] gave an upper bound for the runtime of  $O((\frac{1}{k} + \frac{1}{\lambda})n \log n + (k + \lambda)n)$ . From this and some experiments, the suggestion to take  $k = \lambda$  was derived, reducing the parameter space to the single parameter  $\lambda$ . Since only an upper bound for the runtime was used to obtain this suggestion, again this is an intuitive argument, but not a rigorous one.

For the parameter settings  $p = \lambda/n$ ,  $c = 1/\lambda$ , and arbitrary  $\lambda$  a more precise runtime analysis [DD15b], again on the ONEMAX test function class, gave a tight order of magnitude for the expected runtime of

$$\Theta \left( \max \left\{ \frac{n \log(n)}{\lambda}, \frac{n \lambda \log \log(\lambda)}{\log(\lambda)} \right\} \right),$$

which is minimized exactly by the parameter choice  $\lambda = \Theta(\sqrt{\log(n) \log \log(n) / \log \log \log(n)})$ . As said above, we shall prove that also all other choice of mutation probability, crossover bias, and offspring population size lead to this or a worse runtime.

## 3 Runtime Analysis

*Runtime analysis* is one of the most successful theoretical tools to understand the performance of evolutionary algorithms. The *runtime* or *optimization time* of an algorithm (e.g., our  $(1 + (\lambda, \lambda))$  GA) on a problem instance (e.g., the ONEMAX function) is the number of fitness evaluations that are performed until for the first time an optimal solution is evaluated.

If the algorithm is randomized (like our  $(1 + (\lambda, \lambda))$  GA), this is a random variable  $T$ , and we usually make statements on the expected value  $E[T]$  or give bounds that hold with some high probability, e.g.,  $1 - 1/n$ . When regarding a problem with more than one instance (e.g., traveling salesman instance on  $n$  cities), we take a worst-case view. This is, we regard the maximum expected runtime over all instances, or we make statements like that the runtime satisfies a certain bound for all instances.

In this work, the optimization problem we regard is the classic ONEMAX test problem consisting of the single instance  $\text{OM} : \{0, 1\}^n \rightarrow \{0, 1, \dots, n\}; x \mapsto \sum_{i=1}^n x_i$ , that is, maximizing the number of ones in a bit-string. Despite the simplicity of the ONEMAX problem, analyzing randomized search heuristics on this function has spurred much of the

progress in the theory of evolutionary computation in the last 20 years, as is documented, e.g., in the recent textbook [Jan13].

Of course, when regarding the performance on a single test instance, then we should ensure that the algorithm does not exploit the fact that there is only one instance. A counter-example would be the algorithm that simply evaluates and outputs  $x^* = (1, \dots, 1)$ , giving a perfect runtime of 1. One way of ensuring this is that we restrict ourselves to unbiased algorithms (see [LW12]) which treat bit-positions and bit-values in a symmetric fashion. Consequently, an unbiased algorithm for the ONEMAX problem has the same performance on all problems with isomorphic fitness landscape, in particular, on all (generalized) ONEMAX functions  $\text{OM}_z : \{0, 1\}^n \rightarrow \{0, 1, \dots, n\}; x \mapsto \text{eq}(x, z)$  for  $z \in \{0, 1\}^n$ , where  $\text{eq}(x, z)$  denotes the number of bit-positions in which  $x$  and  $z$  agree. It is easy to see that the  $(1 + (\lambda, \lambda))$  GA is unbiased (for all parameter settings).

## 4 Notation and Technical Tools

In this section, besides fixing some very elementary notation, we collect the main technical tools we shall use. Mostly, these are large deviations bounds of various types. For the convenience of the reader, we first state the known ones. We then prove a tail bound for sums of geometric random variables with expectations bounded from above by the reciprocals of the first positive integers. We finally state the well-known additive drift theorem.

### 4.1 Notation

We write  $[a..b]$  to denote the set  $\{z \in \mathbb{Z} \mid a \leq z \leq b\}$  of integers between  $a$  and  $b$ . We write  $\log(n)$  to denote the binary logarithm of  $n$  and  $\ln(n)$  to denote the natural logarithm of  $n$ . However, to avoid unnecessary case distinctions when taking iterated logarithms, we define  $\log(n) := 1$  for all  $n \leq 2$  and  $\ln(n) := 1$  for all  $n \leq e$ . For the readers' convenience, we now collect some tools from probability theory which we will use regularly.

We occasionally need the expected value of a binomially distributed random variable  $X \sim \mathcal{B}(n, p)$  conditional on that the variable has at least a certain value  $k$ . An intuitive (but wrong) solution to this question is that this  $E[X \mid X \geq k]$  should be around  $k + p(n - k)$ , because we know already that at least  $k$  of the  $n$  independent trials are successes and the remaining  $(n - k)$  trials still have their independent success probability of  $p$ . While this argument is wrong, an upper bound of this type can be shown by elementary means. Since we have not seen this made explicit in the EA literature, we shall also give the short proof.

**Lemma 1.** *Let  $X$  be a random variable with binomial distribution with parameters  $n$  and  $p \in [0, 1]$ . Let  $k \in [0..n]$ . Then*

$$E[X \mid X \geq k] \leq k + (n - k)p \leq k + E[X].$$

*Proof.* Let  $X_1, \dots, X_n$  be independent binary random variables with  $\Pr[X_i = 1] = p$  for all  $i \in [1..n]$ . Then  $X = \sum_{i=1}^n X_i$  has a binomial distribution with parameters  $n$  and  $p$ . Conditioning on  $X \geq k$ , let  $\ell := \min\{i \in [1..n] \mid \sum_{j=1}^i X_j = k\}$ . Then

$E[X \mid X \geq k] = \sum_{i=1}^n \Pr[\ell = i \mid X \geq k] E[X \mid \ell = i]$ . Note that  $\ell \geq k$  by definition. Note also that  $(X \mid \ell = i) = k + \sum_{j=i+1}^n X_j$  with unconditioned  $X_j$ . In particular,  $E[X \mid \ell = i] = k + (n - i)p$ . Consequently,  $E[X \mid X \geq k] = \sum_{i=1}^n \Pr[\ell = i \mid X \geq k] E[X \mid \ell = i] \leq \sum_{i=k+1}^n \Pr[\ell = i \mid X \geq k] (k + (n - k)p) = k + (n - k)p$ .  $\square$

Also, we shall use the following well-known fact.

**Lemma 2.** *Let  $X$  be a non-negative integral random variable. Then  $E[X] = \sum_{i=1}^{\infty} \Pr[X \geq i]$ .*

## 4.2 Known Chernoff Bounds

The following *large deviation bounds* are well-known and can be found, e.g., in [Doe11]. We call all these bounds Chernoff bounds despite the fact that it is now known that some have been found earlier by other researchers.

**Theorem 3** (Classic Chernoff bounds). *Let  $X_1, \dots, X_n$  be independent random variables taking values in  $[0, 1]$ . Let  $X = \sum_{i=1}^n X_i$ .*

- (a). *Let  $\delta \geq 0$ . Then  $\Pr[X \geq (1 + \delta)E[X]] \leq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^{E[X]}$ .*
- (b). *Let  $\delta \in [0, 1]$ . Then  $\Pr[X \geq (1 + \delta)E[X]] \leq \exp(-\delta^2 E[X]/3)$ .*
- (c). *Let  $d \geq 6E[X]$ . Then  $\Pr[X \geq d] \leq 2^{-d}$ .*
- (d). *Let  $\delta \in [0, 1]$ . Then  $\Pr[X \leq (1 - \delta)E[X]] \leq \exp(-\delta^2 E[X]/2)$ .*
- (e). *Let  $X_1, \dots, X_n$  be independent random variables each taking values in some interval of length at most one. Let  $X = \sum_{i=1}^n X_i$ . Let  $\lambda \geq 0$ . Then  $\Pr[X \leq E[X] - \lambda] \leq \exp(-2\lambda^2/n)$  and  $\Pr[X \geq E[X] + \lambda] \leq \exp(-2\lambda^2/n)$ .*

Chernoff bounds also hold for *hypergeometric* distributions. Let  $A$  be any set of  $n$  elements. Let  $B$  be a subset of  $A$  having  $m$  elements. If  $Y$  is a random subset of  $A$  of  $N$  elements (chosen uniformly at random from all  $N$ -element subsets of  $A$ , then  $X := |Y \cap B|$  has a hypergeometric distribution with parameters  $(n, N, m)$ .

**Theorem 4** (Chernoff bounds for hypergeometric distributions). *If  $X$  has a hypergeometric distribution with parameters  $(n, N, m)$ , then  $E[X] = Nm/n$  and  $X$  satisfies all Chernoff bounds given in Theorem 3.*

## 4.3 Drift Analysis

*Drift analysis* comprises a couple of methods to derive from information about the expected progress (e.g., in terms of the fitness distance) a result about the time needed to achieve a goal (e.g., finding an optimal solution). We shall several times use the following *additive drift* theorem from [HY01] (see also Theorem 2.7 in [OY11]).

**Theorem 5** (additive drift theorem). *Let  $X_0, X_1, \dots$  be a sequence of random variables taking values in a finite set  $S \subseteq \mathbb{R}_{\geq 0}$ . Let  $T := \min\{t \geq 0 \mid X_t = 0\}$ . Let  $\delta > 0$ .*



(i) If for all  $t$ , we have  $E[X_t - X_{t+1} | X_t > 0] \geq \delta$ , then  $E[T | X_0] \leq X_0/\delta$ .

(ii) If for all  $t$ , we have  $E[X_t - X_{t+1} | X_t > 0] \leq \delta$ , then  $E[T | X_0] \geq X_0/\delta$ .

In many situation, the progress  $X_t - X_{t+1}$  is stronger when the process is far from the target, that is, when  $X_t$  is large. A particular, but seemingly very common special case is that the progress is indeed proportional to  $X_t$ . Such a situation is called *multiplicative drift*. Drift theorems giving upper bounds for the hitting time were given in [DJW12] and [DG13]. Transforming upper bounds on a multiplicative progress into good lower bounds for hitting times requires additional assumptions. Witt gives the following very useful theorem (Theorem 2.2 in [Wit13]).

**Theorem 6** (multiplicative drift, lower bound). *Let  $S \subset \mathbb{R}$  be a finite set of positive numbers with minimum 1. Let  $X_0, X_1, \dots$  be a sequence of random variables over  $S$  such that  $X_t \geq X_{t+1}$  for any  $t \geq 0$ . Let  $s_{\min} > 0$ . Let  $T$  be the random variable that gives the first point in time  $t \geq 0$  for which  $X_t \leq s_{\min}$ . If there exist positive reals  $\beta, \delta \leq 1$  such that, for all  $s > s_{\min}$  and all  $t \geq 0$  with  $\Pr[X_t = s] > 0$ ,*

$$(1) \ E[X_t - X_{t+1} | X_t = s] \leq \delta s,$$

$$(2) \ \Pr[X_t - X_{t+1} \geq \beta s | X_t = s] \leq \beta \delta / \ln(s),$$

*then for all  $S_0 \in S$  with  $\Pr[X_0 = s_0] > 0$ , we have  $E[T | X_0 = s_0] \geq \frac{\ln(s_0) - \ln(s_{\min})}{\delta} \cdot \frac{1-\beta}{1+\beta}$ .*

## 5 Main Result and Proof

As described in Section 2, a combination of intuitive considerations and rigorous work made [DDE13, DD15b] suggest the parameter choice  $\lambda = \lambda^* := \sqrt{\frac{\log(n) \log \log(n)}{\log \log \log(n)}}$ ,  $p^* = \lambda^*/n$ , and  $c^* = 1/\lambda^*$  for the optimization of the ONEMAX test function class, yielding an expected optimization time of  $F^* = \frac{n \log n}{\lambda^*} = n \sqrt{\frac{\log(n) \log \log \log(n)}{\log \log(n)}}$ . It was also proven that with  $p$  and  $c$  functionally depending on  $\lambda$  as above,  $\lambda = \Theta(\lambda^*)$  is the optimal choice and the only optimal choice.

In this section, we complete this picture by proving rigorously that no combination of the parameters  $p$ ,  $c$ , and  $\lambda$ , all possibly depending on  $n$ , can lead to an expected optimization time of asymptotic order strictly better than  $F^*$ . We also show that not many parameter combinations can give this optimal expected runtime.

**Theorem 7.** *Let  $\lambda^* := \sqrt{\frac{\log(n) \log \log(n)}{\log \log \log(n)}}$  and  $F^* = \frac{n \log n}{\lambda^*} = n \sqrt{\frac{\log(n) \log \log \log(n)}{\log \log(n)}}$ .*

- *For arbitrary parameters  $\lambda \in [0..n]$ ,  $p \in [0, 1]$  and  $c \in [0, 1]$ , all being functions on  $n$ , the  $(1 + (\lambda, \lambda))$  GA has an expected optimization time of  $E[F] = \Omega(F^*)$ .*
- *If some parameter combination  $(\lambda, p, c)$  leads to an expected optimization time of  $E[F] = \Theta(F^*)$ , then*
  - $\lambda = \Theta(\lambda^*)$ ,
  - $p = \Omega(\lambda^*/n)$  and  $p = (1/n) \exp(O(\sqrt{\log(n) \log \log \log(n) / \log \log(n)}))$ , and

$$- c = \Theta(1/pn).$$

We remark that *the same lower bound holds for the natural modification of the  $(1 + (\lambda, \lambda))$  GA in which the best of all mutation and crossover offspring competes in the final selection step with the parent individual* (and not only the best crossover offspring). The proofs below are written up in a way that this is easy to check, but to keep the paper readable we do not explicitly formulate all statements for both version of the algorithm. Consequently, for the ONEMAX testfunction, this modification does not give an asymptotic runtime improvement. In a practical application, however, there is no reason to not exploit possible exceptionally good mutation offspring. So here this modification seems very advisable.

To ease the presentation, we shall always parameterize these values by  $p = k/n$  and  $c = r/k$  for some  $k \in [0, n]$  and  $r \in [0, k]$  (hence  $k$  and  $r$  may also depend on  $n$ ). In this language, the previously suggested values are  $k^* = \lambda^*$  and  $r^* = 1$ , and the main result of this work is that

- (i) no parameter setting gives a better expected optimization time than the  $\Theta(F^*)$  stemming from these parameters, and
- (ii) any parameter tuple  $(\lambda, k, r)$  that leads to an asymptotic optimization time of  $\Theta(F^*)$  satisfies  $\lambda = \Theta(\lambda^*)$ ,  $k = \Omega(k^*)$  and  $k = \exp(O(\sqrt{\log(n) \log \log \log(n) / \log \log(n)}))$ , and  $r = \Theta(r^*)$ .

A side remark: Another implicit parameter choice done in [DDE13] is to use the same offspring population size  $\lambda$  for the mutation phase and the crossover phase. One could well imagine having different numbers  $\lambda_m$  and  $\lambda_c$  of offspring for both phases. This may make sense in practical applications or when performing a theoretical analysis that takes care of constant factors. In this work, where we are only precise up to the asymptotic order of magnitude, the optimization time is of asymptotic order equal to the product of the number of iterations and  $\max\{\lambda_m, \lambda_c\}$ . Hence, unless one believes that a smaller offspring population size can reduce the number of iterations (which is not what our proofs suggest), there is for us no use of not taking both offspring population sizes equal to  $\max\{\lambda_m, \lambda_c\}$ .

## 5.1 Overview of the Proof

Given apparent difficulty (see [DD15b]) of determining the runtime of the  $(1 + (\lambda, \lambda))$  GA already for settings  $k = \lambda$  and  $r = 1$  suggested in [DDE13], the common approach of determining the optimal parameter settings by conducting a precise runtime analysis for all parameter combinations  $(\lambda, k, r)$  seems not very promising. Therefore, shall rather analyze particular parts of the optimization process in detail and from these extract necessary conditions for the parameters to allow an expected optimization time of order  $O(F^*)$ . To make it more visible how the different arguments work together, let us start with a brief overview of the analysis.

Let a tuple  $(\lambda, p = k/n, c = r/k)$  as described above be given. We denote by  $T$  the number of iterations the  $(1 + (\lambda, \lambda))$  GA with these parameters performs until an optimal solution is found (we have  $T = 0$  if the random initial search point is already optimal). We denote by  $F$  the optimization time of this  $(1 + (\lambda, \lambda))$  GA, that is, the



number of fitness evaluations performed until an optimal solution is evaluated. This is one if the random initial search point was optimal. We roughly have  $F \approx 2\lambda T$ , but see Proposition 10 and the text around it for the details.

We say that a tuple of parameters is *optimal* if the resulting optimization time is  $O(F^*)$ . This is, for the moment, a slight abuse of language, but as this section will show, these are indeed the parameters that lead to the asymptotically optimal runtime, since (as we will see) no better runtime than  $\Omega(F^*)$  can be achieved with any parameter setting. The proof of the Theorem 7 then consists of the following arguments, which all can be shown independent of the others. Since we aim at an asymptotic result only, we can freely assume that  $n$  is sufficiently large.

- In Lemma 11, we make the elementary observation that  $E[F] \geq \min\{\lambda, 2^n\}/2$ . Consequently,  $\lambda \leq 2F^*$  in any optimal parameter set.
- In Lemma 12, we show that

$$E[F] = \min\{\Omega(r^{-1} \exp(\Theta(r))n \log n), \exp(\Omega(r))n^2 \log n, \exp(\Omega(n^{1/16}))\}$$

when  $k \geq \sqrt{n}$  and  $\lambda = \exp(o(n^{1/16}))$ . Since this runtime is at least  $\Omega(n \log n)$ , together with the previous item (showing that  $\lambda$  cannot be too large), we obtain that  $k \leq \sqrt{n}$  in an optimal parameter set.

- In Lemma 13, we show that for  $0 < k \leq n/12$ , we have  $E[F] = \Omega(\frac{n \log n}{k})$ . Hence  $k = \Omega(\lambda^*)$  in an optimal parameter setting.
- In Lemma 14, we show that when  $\omega(1) = k \leq \sqrt{n}$ , then  $E[F] = \Omega(n \log n \min\{\frac{\exp(\Omega(r))}{\lambda r}, \frac{n^3}{\lambda}, \frac{\exp(\Omega(k))}{k}\})$ . Since we know already that  $\lambda \leq n^3$  and  $k = \omega(1)$  in an optimal parameter setting, this result implies that an optimal parameter set has  $\lambda = \Omega(\lambda^* \exp(\Omega(r))/r)$ .
- In Lemma 15, we show  $E[F] = \Omega(n\lambda/k)$  when  $k \leq n/4$  (which we know already). Consequently, in an optimal set of parameters  $\lambda$  cannot be excessively large, e.g.,  $\lambda \leq \exp(k/120)$ .
- In Lemma 16, we show that if  $k \leq n/80$ ,  $\lambda \leq \exp(k/120)$ ,  $\lambda = \exp(o(n))$ , and  $\lambda = \omega(1)$ —all of this holds in an optimal parameter setting as shown above—then  $E[F] = \Omega(\frac{n\lambda \log \log(\lambda)}{r \log \lambda})$ . This result together with Lemma 14 implies that the optimal runtime is  $\Theta(F^*)$  and that we have  $\lambda = \Theta(\lambda^*)$  and  $r = \Theta(1)$  in an optimal parameter setting.

This shows the main claim of this work, namely that  $F^*$  is asymptotically the best runtime one can achieve with a clever choice of all parameters of the  $(1 + (\lambda, \lambda))$  GA. The above also shows that an optimal parameter set has  $\lambda = \Theta(\lambda^*)$  and  $r = \Theta(1)$ . For the mutation probability, the above only yields  $k = \Omega(\lambda^*)$  and  $k = O(\sqrt{n})$ . In Lemma 17, we show that  $k = \exp(O(\sqrt{\log(n) \log \log \log(n) / \log \log(n)}))$  is a necessary condition for having a  $\Theta(F^*)$  runtime.

We do not know if the interval of optimal  $k$  values can be further reduced. An inspection of the upper bound proof in [DD15b] suggests that, with more effort than

there, also slightly larger  $k$ -values than  $\Theta(\lambda^*)$  (together with  $\lambda = \Theta(\lambda^*)$  and  $r = \Theta(1)$ ) could lead to the optimal expected runtime of  $\Theta(F^*)$ . We do not follow up on this question, because we do not feel that it justifies the effort of extending the technical proof of [DD15b]. It is quite clear that there is no algorithmic advantage of using a larger than necessary  $k$ -value. The main (unfavorable) difference would be that than an efficient implementation of the mutation operator in expected time  $\Theta(k)$  would have an increased complexity.

We face two main difficulties in this proof. One are the apparent dependencies introduced by the two intermediate selection steps and the fact that all mutation offspring have the same Hamming distance from the parent. That the latter creates additional challenges can be easily seen in the lengthy proof of Lemma 13, which simply tries to use the classic argument that one needs at least a total number of  $\Theta(n \log n)$  bit-flips to make sure that each initially incorrect bit was flipped at least once.

The second difficulty is that even parameter combinations that are far from those leading to the optimal runtime can lead to runtimes very close to the optimal one. An example (given here without proof) is that for say  $k = \sqrt{n}$  and  $\lambda = \lambda^*$  and  $r = 1$ , the optimization process strongly resembles the one of the  $(1 + \lambda)$  EA with  $\lambda$  below the cut-off point. Consequently, the  $(1 + (\lambda, \lambda))$  GA for these parameters has an optimization time of  $\Theta(n \log n)$ , which is relatively close to  $F^*$  given uncommonly large mutation probability.

## 5.2 Proofs

This this longer subsection, we prove the results outlined above. We frequently use the following notation. For  $x \in \{0, 1\}^n$ , we call  $d(x) := n - \text{OM}(x)$  its *fitness distance*. Let  $x, x', y \in \{0, 1\}^n$ . Then

$$g(x, x') := |\{i \in [1..n] \mid x_i = 0 \wedge x'_i = 1\}|$$

is the number of *good bits of  $x'$  (with respect to  $x$ )*. Analogously,

$$b(x, x') := |\{i \in [1..n] \mid x_i = 1 \wedge x'_i = 0\}|$$

is the number of *bad bits of  $x'$  (with respect to  $x$ )*. Note that, trivially,  $g(x, x') + b(x, x') = H(x, x')$ , the Hamming distance of  $x$  and  $x'$ . Similarly, we define “the number of good bits of  $x'$  that made it into  $y$ ” and “the number of bad bits of  $x'$  that made it into  $y$ ” by

$$\begin{aligned} g(x, x', y) &:= |\{i \in [1..n] \mid x_i = 0 \wedge x'_i = 1 \wedge y_i = 1\}|, \\ b(x, x', y) &:= |\{i \in [1..n] \mid x_i = 1 \wedge x'_i = 0 \wedge y_i = 0\}|. \end{aligned}$$

In the following, we always assume that we consider a run of the  $(1 + (\lambda, \lambda))$  GA with general parameter setting  $\lambda$ ,  $p = k/n$ , and  $c = r/k$ , which may all depend on the problem size  $n$ . Since we are interested in an asymptotic result, we may assume that  $n$  is sufficiently large. We use the variables of the algorithm description, e.g.,  $x$ ,  $x^{(i)}$ ,  $x'$ , etc. without further explicit reference to the algorithm.

We now prove the ingredients forming the proof of the main result. We prove these results not only for the minimal parameter range needed in the proof of the main result, but rather for those ranges where the main arguments work well. At the same time, we

do not aim at the absolutely widest parameter range and we occasionally do not aim at the sharpest possible bound if this would significantly increase the proof complexity. We aim at keeping the proofs of the partial results independent, both to ease reading and to allow an easier understanding how the main proof decomposes into the partial results. For this reason, all of the following lemmas are proven independently apart from possibly relying on the two elementary propositions 8 and 10.

The first of these proposition is a technical tool showing that extraordinarily large fitness gains occurs rarely. This allows in the following to assume that the algorithm indeed once has a parent individual  $x$  with roughly a certain fitness.

**Proposition 8.** *Let  $x$  be a search point with  $d := d(x)$  satisfying  $d \leq 0.6n$ . Then the probability that one iteration of the  $(1 + (\lambda, \lambda))$  GA with arbitrary parameter settings creates a search point  $y$  with  $d(y) \leq d/2$ , is  $\lambda(\lambda + 1) \exp(-\Omega(d))$ .*

To prove this proposition, we need the elementary fact that standard bit mutation hardly reduces  $d(\cdot)$  by 50% or more.

**Proposition 9.** *Let  $p \in [0, 1]$ ,  $x \in \{0, 1\}^n$  with  $d := d(x) \leq 0.6n$ , and  $y$  be obtained from flipping each bit of  $x$  independently with probability  $p$ . Then  $\Pr[d(y) \leq 0.5d] = \exp(-\Omega(d))$ .*

*Proof.* Let first  $0.1n \leq d \leq 0.6n$ . Then  $E[d(y)] \geq \min\{d, 0.4n\}$  regardless of  $p$ . Consequently,  $\Pr[d(y) \leq d/2] \leq \Pr[d(y) \leq E[d(y)] - 0.05n] \leq \exp(-\Theta(n))$  by the additive Chernoff bound (Theorem 3 (e)).

Let now  $d \leq 0.1n$ . Let  $g := g(x, y)$  and  $b = b(x, y)$ . Trivially, we have  $d(y) = d - g + b$ . Let first  $p \leq 1/4$ . Since  $g$  is binomially distributed with parameters  $d$  and  $p$ , we have  $E[g] = dp \leq d/4$  and  $\Pr[g \geq d/2] \leq \exp(-\Omega(d))$  by the multiplicative Chernoff bound (Theorem 3 (d)). We thus have  $\Pr[d(y) \leq d/2] \leq \Pr[g \geq d/2] \leq \exp(-\Omega(d))$ . Let now  $p \geq 1/4$ . Then  $E[b] = (n - d)p \geq 0.225n$  and  $\Pr[b \leq 0.1n] \leq \exp(-\Omega(n))$ . Since trivially  $g \leq d \leq 0.1n$ , we have  $\Pr[d(y) \leq d/2] \leq \Pr[b \leq 0.1n] \leq \exp(-\Omega(n))$ .  $\square$

*Proof of Proposition 8.* To ease the calculations, we use the following Gedankenexperiment. Imagine that the  $(1 + (\lambda, \lambda))$  GA does not select a winning individual  $x'$  at the end of the mutation phase, but instead creates  $\lambda$  crossover offspring from each of the  $\lambda$  mutation offspring. Clearly, the set of  $\lambda$  crossover offspring from a true run of the algorithm is contained in this set of  $\lambda^2$  offspring. Hence it suffices to show that none of the  $\lambda^2$  offspring from the Gedankenexperiment and none of the  $\lambda$  mutation offspring has a fitness distance of  $d/2$  or better.

Let  $\tilde{y}$  be a crossover offspring of the Gedankenexperiment. Let  $\tilde{x}$  be the mutation offspring that was used in the crossover giving rise to  $\tilde{y}$ . Then  $\tilde{x}$  is obtained from  $x$  by flipping each bit independently with probability  $k/n$ —the  $(1 + (\lambda, \lambda))$  GA creates  $\tilde{x}$  algorithmically different, namely by first sampling  $\ell$  and then flipping  $\ell$  bits, but the result is that  $\tilde{x}$  has the distribution described above due to the choice of  $\ell$ . Now  $\tilde{y}$  is obtained from a biased crossover of  $x$  and  $\tilde{x}$ . Since each bit of  $\tilde{x}$  makes it into  $\tilde{y}$  only with probability  $r/k$ , we see that we have  $\tilde{y}_i \neq x_i$  with probability  $(k/n) \cdot (r/k) = r/n$  independently for all  $i \in [1..n]$ . Consequently,  $\tilde{y}$  has the same distribution as if it was generated from  $x$  by standard bit mutation with mutation rate  $r/n$ .

Since all mutation and crossover offspring are distributed as if generated via standard bit mutation (with some mutation rate that does not matter here), Proposition 9 and a simple union bound over the  $\lambda(\lambda + 1)$  mutation and crossover offspring shows that with probability at least  $1 - \lambda(\lambda + 1) \exp(-\Omega(d))$  none of these has a fitness distance of  $d/2$  or better.  $\square$

The second proposition shows that, apart from exceptional cases, we can freely switch between the number of iterations  $T$  and the number of fitness evaluations  $F$  needed to find an optimum. This is a well-known fact, so we give its proof merely for reasons of completeness. Recall that the optimization time is defined to be the number of fitness evaluations until for the first time an optimal solution is evaluated. Consequently, if say the first mutation offspring by chance is an optimal solution, then the optimization time  $F$  would be 2. The number of iterations  $T$ , though, would be 1, so the estimate  $F = \Omega(\lambda T)$  is not valid. The following lemma shows this exceptional case only occurs for  $E[T] < 2$ , so that usually we can (and will without further notice) use the argument  $E[F] = \Omega(\lambda E[T])$ .

**Proposition 10.** *If  $E[T] \geq 2$ , then  $E[F] = \Theta(\lambda E[T])$ .*

*Proof.* By definition of  $F$  and  $T$ , we have  $T = \lceil (F - 1)/2\lambda \rceil \leq (F - 1)/2\lambda + 1$ . Consequently,  $F \geq 2(T - 1)\lambda + 1$  and  $E[F] \geq E[2(T - 1)\lambda + 1] \geq 2(E[T] - 1)\lambda \geq E[T]\lambda$  when  $E[T] \geq 2$ . Since  $F \leq 2\lambda T + 1$ , we also have  $E[F] = O(\lambda E[T])$ .  $\square$

We now start proving a number of lower bounds for the runtime of the  $(1 + (\lambda, \lambda))$  GA. They do not logically rely on each other. The first result shows that, unless  $\lambda$  is excessively large, the expected optimization time is at least  $\Omega(\lambda)$ .

**Lemma 11.**  $E[F] \geq \min\{\lambda, 2^n\}/2$ .

*Proof.* The proof builds on the following simple observation: Let  $\tilde{x}$  be a mutation offspring generated in the first iteration. Then  $\tilde{x}$  is uniformly distributed in  $\{0, 1\}^n$ . Indeed, let  $x$  be the random initial search point, which is uniformly distributed in  $\{0, 1\}^n$ , which is equivalent to saying that each  $x_i$  independently is equal to 1 with probability  $1/2$  (and is equal to 0 otherwise). Now  $\tilde{x}$  is generated from  $x$  by flipping each bit independently with probability  $k/n$ . Consequently, the bits of  $\tilde{x}$  are independent. We also compute  $\Pr[\tilde{x}_i = 1] = \Pr[x_i = 0](k/n) + \Pr[x_i = 1](1 - k/n) = 1/2$ . Hence  $\tilde{x}$  is uniformly distributed in  $\{0, 1\}^n$ .

With this preliminary consideration, the proof of the lemma is very easy. Let  $L$  be a non-negative integer. Let  $x_0, x_1, \dots, x_L$  be the initial random search point and the first  $L$  mutation offspring. Note that each of these search points individually is uniformly distributed in  $\{0, 1\}^n$ . Consequently, by a simple union bound, the probability that one of these search points is the optimum is at most  $(L + 1)2^{-n}$ . In other words, the number  $F$  of fitness evaluations until an optimal solution is found, satisfies  $\Pr[F \geq L + 2] \geq 1 - (L + 1)2^{-n}$  for all  $0 \leq L \leq \lambda$ . By Lemma 2, taking  $K = \min\{\lambda + 1, 2^n\}$ , we compute  $E[F] = \sum_{i=1}^{\infty} \Pr[F \geq i] \geq \sum_{i=1}^K \Pr[F \geq i] \geq \sum_{i=1}^K (1 - (i - 1)2^{-n}) = K - \frac{K(K-1)}{2}2^{-n} = K(1 - 2^{-n-1}(K - 1)) \geq \min\{\lambda, 2^n\}/2$ .  $\square$

We proceed by regarding the case that  $k$  is large, say  $k \geq \sqrt{n}$ . While this is much larger than all values of  $k$  that lead to the optimal expected runtime, the proof is not very

simple. The reason is that even such large values for  $k$  can give a near-optimal runtime of  $O(n \log n)$  for suitable choices of the other parameters, e.g., small values for  $\lambda$  and  $r = 1$  (we do not prove this statement).

**Lemma 12.** *If  $k \geq \sqrt{n}$  and  $\lambda = \exp(o(n^{1/16}))$ , then*

$$E[F] = \min\{\Omega(r^{-1} \exp(\Theta(r))n \log n), \exp(\Omega(r))n^2 \log n, \exp(\Omega(n^{1/16}))\},$$

*which attains its asymptotically optimal value  $\Omega(n \log n)$  for  $r = \Theta(1)$ .*

*Proof.* We start by analyzing the progress the  $(1 + (\lambda, \lambda))$  GA makes in one iteration starting with a search point  $x$  having fitness distance  $d := d(x) \in [n^{3/4}, n^{7/8}]$ . More precisely, denote by  $z$  an individual with maximal fitness among all mutation and crossover offspring generated in this iteration and among the parent  $x$ . Needless to say,  $z$  can be the parent  $x$ , the crossover winner  $y$ , or the mutation winner  $x'$ . To use drift analysis, we shall regard the progress  $d(x) - d(z)$ . Note that this is 0 if  $\max\{f(x'), f(y)\} \leq f(x)$ . Note also that  $d(x) - d(z) = f(z) - f(x)$ .

Let  $\tilde{x}$  be a mutation offspring. Let  $\tilde{g} = g(x, \tilde{x})$  be the number of good bits of  $\tilde{x}$ . Since  $\tilde{g}$  follows a binomial distribution with parameters  $d$  and  $k/n$ , we have  $E[\tilde{g}] = dk/n \geq n^{1/4}$  and  $\Pr[\tilde{g} \geq 2dk/n] \leq \exp(-(dk/n)/3) \leq \exp(-\Omega(n^{1/4}))$ . Hence only with probability at most  $\lambda \exp(-\Omega(n^{1/4}))$ , there is a mutation offspring with at least  $2dk/n$  good bits; in this rare case we estimate the progress  $f(z) - f(x)$  via the trivial bound  $f(z) - f(x) \leq n$ . Similarly, in the exceptional case that  $\ell < k/2$ , which occurs with probability at most  $\exp(-\Omega(k)) \leq \exp(-\Omega(n^{1/2}))$ , we again estimate  $f(z) - f(x) \leq n$ .

Hence let us now analyze the progress in the *regular situation* that no mutation offspring has  $2dk/n$  good bits or more (and thus  $g(x, x') < 2dk/n$ ) and that  $\ell \geq k/2$  (and thus  $x'$  has at least  $b(x, x') \geq \ell - (2dk/n) \geq (k/2) - (2dk/n) = k((1/2) - 2n^{-1/8}) \geq k/4$  bad bits). Since  $b(x, x') > g(x, x')$ , we have  $z \neq x'$ , so it remains to analyze the crossover offspring. Consider an offspring  $\tilde{y}$  generated in the crossover phase.

Let us consider first *the case that  $r \geq n^{1/16}$* . Then  $\tilde{b} := b(x, x', \tilde{y})$  satisfies  $E[\tilde{b}] \geq (k/4) \cdot (r/k) = r/4$ . Hence with probability  $1 - \exp(-\Omega(r)) \geq 1 - \exp(-\Omega(n^{1/16}))$ , the crossover offspring  $\tilde{y}$  has taken at least  $k/8$  bad bits from  $x'$ . This is more than the number of good bits  $x'$  has, so regardless of how many good bits make it into  $\tilde{y}$ , we have  $f(\tilde{y}) \leq f(x)$ . Consequently, with probability  $1 - \lambda \exp(-\Omega(n^{1/16}))$ , no crossover offspring has a fitness better than  $x$ , and hence  $f(z) = f(x)$ . For the remaining probability  $\lambda \exp(-\Omega(n^{1/16}))$ , we estimate  $f(z) - f(x) \leq n$ . In total, if  $r \geq n^{1/16}$ , we have  $E[f(z) - f(x)] \leq n\lambda \exp(-\Omega(n^{1/16})) = \lambda \exp(-\Omega(n^{1/16}))$ .

We now turn to *the case that  $r < n^{1/16}$* . In this case,  $\tilde{g} := g(x, x', \tilde{y})$  satisfies  $E[\tilde{g}] \leq (2dk/n) \cdot (r/k) = 2dr/n \leq 2n^{-1/16}$ . We regard separately the situations that  $\tilde{g} = 0$ ,  $\tilde{g} \in [1..47]$ ,  $\tilde{g} \in [48..E[\tilde{b}]/2]$ , and  $\tilde{g} \geq E[\tilde{b}]/2$ . Clearly, when  $\tilde{g} = 0$ , we have  $f(\tilde{y}) \leq f(x)$ . Markov's inequality shows that good bits exist only with probability  $E[\tilde{g}] \leq 2dr/n$ , hence,  $\Pr[\tilde{g} \in [1..47]] \leq \Pr[\tilde{g} \geq 1] \leq 2dr/n$ . Conditioning on  $\tilde{y}$  having between one and 47 good bits, we trivially observe  $f(\tilde{y}) - f(x) \leq 47$ . However, for  $\tilde{y}$  to have a fitness better than  $f(x)$ , it is necessary (but not sufficient) that at most 46 bad bits are copied from  $x'$  to  $\tilde{y}$ . The probability of this event, which is independent of any event regarding good bits only, is at most  $\exp(-\Omega(E[\tilde{b}])) \leq \exp(-\Theta(r))$ , because the expected number  $E[\tilde{b}]$  of bad bits copied into  $\tilde{y}$  is  $\Theta(r)$ . By Theorem 3 (a), the probability that 48 or more good bits

are copied into  $\tilde{y}$  is  $O(n^{-3})$ , hence  $\Pr[\tilde{g} \in [48.. \lfloor E[\tilde{b}]/2 \rfloor]] \leq \Pr[\tilde{g} \geq 48] = O(n^{-3})$ . In this situation, for  $f(\tilde{y})$  to be larger than  $f(x)$ , we need  $\tilde{b} < \tilde{g} \leq E[\tilde{b}]/2$ , which happens with probability  $\exp(-\Omega(E[\tilde{b}])) \leq \exp(-\Omega(r))$ . Finally, if  $E[\tilde{b}]/2 \geq 48$ , then the probability that  $\tilde{g} \geq E[\tilde{b}]/2$  is at most  $n^{-3-\Omega(E[\tilde{b}])} \leq n^{-3-\Omega(r)}$  by Theorem 3 (a). Hence

$$\begin{aligned} E[\max\{f(\tilde{y}) - f(x), 0\}] &\leq \Pr[\tilde{g} = 0] \cdot 0 \\ &\quad + \Pr[\tilde{g} \in [1..47]] \exp(-\Omega(r)) \cdot 47 \\ &\quad + \Pr[\tilde{g} \in [48.. \lfloor E[\tilde{b}]/2 \rfloor]] \exp(-\Omega(r)) \cdot n \\ &\quad + \Pr[\tilde{g} \geq E[\tilde{b}]/2 \mid E[\tilde{b}]/2 \geq 48] \cdot n \\ &\leq 0 + \frac{2dr}{n} \exp(-\Omega(r)) \cdot 47 + O(n^{-3}) \exp(-\Omega(r)) \cdot n + n^{-3-\Omega(r)} \cdot n \\ &\leq O((\frac{dr}{n} + n^{-2}) \exp(-\Omega(r))). \end{aligned}$$

Since  $y$  is chosen among the crossover offspring  $\tilde{y}$  such that  $f(\tilde{y})$ , and equivalently,  $f(\tilde{y}) - f(x)$  is maximal, we have  $f(y) - f(x) \leq \sum_{\tilde{y}} \max\{f(\tilde{y}) - f(x), 0\}$ , where  $\tilde{y}$  runs over all  $\lambda$  crossover offspring. Consequently,  $E[f(y) - f(x)] = O(\lambda(\frac{dr}{n} + n^{-2}) \exp(-\Omega(r)))$ .

Taking the two cases regarded separately together, we see that for any  $r$  we have  $E[f(z) - f(x)] = E[\max\{0, f(y) - f(x)\}] = \max\{O(\lambda(\frac{dr}{n} + n^{-2}) \exp(-\Omega(r))), \lambda \exp(-\Omega(n^{1/16}))\}$ , when we condition on being in the regular situation. In the general situation, we have  $E[f(z) - f(x)] = \lambda \exp(-\Omega(n^{1/4})n + (1 - \lambda \exp(-\Omega(n^{1/4}))) \max\{O(\lambda(\frac{dr}{n} + n^{-2}) \exp(-\Omega(r))), \lambda \exp(-\Omega(n^{1/16}))\}) = \max\{O(\lambda(\frac{dr}{n} + n^{-2}) \exp(-\Omega(r))), \lambda \exp(-\Omega(n^{1/16}))\}$ . To ease the following multiplicative drift argument, we estimate this bluntly by  $E[f(z) - f(x)] \leq \max\{O(\lambda(\frac{dr}{n} + dn^{-2})) \exp(-\Omega(r)), d\lambda \exp(-\Omega(n^{1/16}))\} = d \max\{O(\lambda \max\{r, n^{-1}\} \exp(-\Omega(r))/n, \lambda \exp(-\Omega(n^{1/16})))\}$ .

Building on this drift statement, we now use Witt's lower bound result for multiplicative drift (Theorem 6). Consider a run of the  $(1 + (\lambda, \lambda))$  GA. For  $t = 0, 1, \dots$ , denote by  $x_t$  the search point  $x$  at the beginning of the  $(t + 1)$ st iteration except if before that once an optimal solution was generated, in this case let  $x_t$  be any optimal solution. By Proposition 8, with probability at least  $1 - \lambda(\lambda + 1) \exp(-\Omega(n^{7/8}))$  the  $(1 + (\lambda, \lambda))$  GA at some time  $t_0$  reaches a search point  $x_{t_0}$  with  $d(x_{t_0}) \in [0.5n^{7/8}, n^{7/8}]$ . We show that in this case, we have an expected optimization time as claimed, which implies that also the unconditioned expectation is of the same order of magnitude.

For  $t = 0, 1, \dots$  define  $X_t = \max\{d(x_{t_0+t}), 1\}$ . Observe that  $X_{t+1} \leq X_t$  for all  $t \geq 0$ . Let  $s_{\min} := n^{3/4}$ . Then we have shown above that if  $X_t = s > s_{\min}$ , then  $E[X_t - X_{t+1}] \leq s \max\{K_1 \lambda \max\{r, n^{-1}\} \exp(-K_2 r)/n, \lambda \exp(-K_3 n^{1/16})\}$  for some absolute constants  $K_1, K_2, K_3$ . Note that the drift of the process  $X_t$  might be smaller than this, because above we took  $z$  as the best individual among parent and all individuals generated in the iteration. The first condition of the drift theorem thus is fulfilled with  $\delta = \max\{K_1 \lambda \max\{r, n^{-1}\} \exp(-K_2 r)/n, \lambda \exp(-K_3 n^{1/16})\}$ . From Proposition 8 we know that  $\Pr[X_{t+1} \leq s/2] \leq \lambda(\lambda + 1) \exp(-\Omega(s)) = \exp(-\Omega(s))$ . Hence for  $n$  (and thus also  $s$ ) sufficiently large, also the second condition of the drift theorem is satisfied (with  $\beta = 1/2$ ); also we have  $E[T] = \Omega(\log n)$  to enable the argument  $E[F] = \Omega(\lambda E[T])$  below. We may thus apply the theorem and derive that the first  $t$  such that  $X_t \leq s_{\min}$  satisfies  $E[t] = \Omega(\frac{\ln(X_0) - \ln(s_{\min})}{\delta}) = \Omega(\min\{\frac{\exp(\Theta(r)n \log n)}{\max\{r, 1/n\}\lambda}, \exp(\Omega(n^{1/16}))/\lambda\})$ . Note that this, naturally, is a lower bound on  $E[T]$ . Consequently,  $E[F] = \Omega(\lambda E[T]) =$



$$\Omega(\min\{\frac{\exp(\Omega(r))}{r}n \log n, \exp(\Omega(r))n^2 \log n, \exp(\Omega(n^{1/16}))\}). \quad \square$$

The following lower bound imitates the classic argument that if in all applications of the mutation operators not enough bits are flipped, then there will be a bit that is initially zero and that was never touched in a mutation operation. The proof is slightly more involved as usual for this type of argument, because our mutation operator uses a hypergeometric distribution.

**Lemma 13.** *Let  $0 < k \leq n/12$  and  $k\lambda = o(n \log n)$ . Let  $\alpha < 1/4$ . Let  $t = \lfloor \alpha n \ln(n)/(k\lambda) \rfloor$ . Then  $\Pr[T \leq t] = \exp(-\Omega(\min\{kt, n^{1-4\alpha}\}))$ . In particular,  $E[F] = \Omega(\frac{n \log n}{k})$ . Consequently, an optimal parameter setting satisfies  $k = \Omega(\sqrt{\log(n) \log \log(n) / \log \log \log(n)}) = \Omega(\lambda^*)$ .*

*Proof.* Using the Chernoff bound of Theorem 3 (d), we see that with probability  $1 - \exp(-\Omega(n))$ , the initial search point has at least  $n/3$  bits valued zero (“missing bits”).

Let us consider what happens in the first  $t = \lfloor \alpha n \ln(n)/(k\lambda) \rfloor$  iterations. Denote by  $\ell_1, \dots, \ell_t$  the values of  $\ell$  chosen by the algorithm in these iterations. Note that the  $\ell_i$  are independent random variables each having a binomial distribution with parameters  $n$  and  $k/n$ . Consequently,  $L := \sum_{i=1}^t \ell_i$  is a sum of  $tn$  independent 0,1 random variables that are one with probability  $k/n$ . Hence we have  $E[L] = tk$ . By the multiplicative Chernoff bound of Theorem 3 (b), we see that with probability  $1 - \exp(-\Omega(tk))$ , we have  $L \leq 2tk$ .

Again exploiting the binomial distribution of the  $\ell_i$ , we derive from Theorem 3 (c) that  $\Pr[\ell_i \geq n/2] \leq 2^{-n/2}$ ; note that here we used that  $k \leq n/12$  and thus  $E[\ell_i] = k \leq n/12$ . Consequently, with probability  $1 - \exp(-\Omega(n))$ , all  $\ell_i$  are at most  $n/2$  (union bound).

In the following, we condition on none of these three rare events occurring. More precisely, we condition on that there are at least  $n/3$  missing bits and we condition on a particular outcome of the  $\ell_i$  that avoids the exceptional events  $L > 2tk$  and  $\ell_i > n/2$  for some  $i \in [1..t]$ . The probability that a particular one of the missing bits is never flipped in the mutation phases of the first  $t$  iterations is

$$\prod_{i=1}^t (1 - \ell_i/n)^\lambda \geq \prod_{i=1}^t \exp(-2\ell_i/n)^\lambda = \exp(-2\lambda L/n) \geq \exp(-4k\lambda t/n) \geq n^{-4\alpha},$$

where we have used in the first step that  $1 - c \geq e^{-2c}$  for  $0 \leq c \leq 1/2$ .

Denote by  $M \subseteq [1..n]$  the set of missing bits and by  $A_i$  the event that bit  $i$  was flipped at least once in some mutation step in the first  $t$  iterations. Then we just showed  $\Pr[A_i] \leq 1 - n^{-4\alpha}$ . We want to show that it is very unlikely that all events  $A_i$  are fulfilled.

Unfortunately, the events  $A_i, i \in M$ , are not independent, since already in a single application of the mutation operator the bits are not treated independently, but according to a hypergeometric distribution. We therefore now show that they satisfy the following negative correlation property:

$$\forall I \subseteq M : \Pr \left[ \bigcap_{i \in I} A_i \right] \leq \prod_{i \in I} \Pr[A_i].$$

We proceed via induction over the cardinality of  $I$ . For  $|I| = 0, 1$ , there is nothing to show. Let  $I \subseteq M$  such that  $|I| \geq 2$ . Let  $j \in I$  and  $I' := I \setminus \{j\}$ . Then

$$\Pr \left[ \bigcap_{i \in I'} A_i \right] = \Pr \left[ \bigcap_{i \in I'} A_i \middle| A_j \right] \Pr[A_j] + \Pr \left[ \bigcap_{i \in I'} A_i \middle| \bar{A}_j \right] \Pr[\bar{A}_j]. \quad (1)$$

It is clear that  $\Pr[\bigcap_{i \in I'} A_i \mid \bar{A}_j]$  is at least as large as  $\Pr[\bigcap_{i \in I'} A_i]$ —conditioning on  $\bar{A}_j$  is equivalent to saying that the random subsets of bits to be flipped are not chosen as subsets of  $[1..n]$ , but of  $[1..n] \setminus \{j\}$ , and this increases the probability of the event  $\bigcap_{i \in I'} A_i$ . More formally, there is the following coupling from the unconditioned probability space into the one conditional on  $\bar{A}_j$ . Whenever in the unconditioned probability space the  $j$ -th bit is flipped in some iteration, we replace this bit-flip by flipping a new bit different from  $j$  and the other bits flipped in this iteration. This is exactly the random experiment done in the probability space conditional on  $\bar{A}_j$ . Clearly, if the event  $\bigcap_{i \in I'} A_i$  holds in the unconditioned space, this is not affected by the coupling. Hence the probability of the event  $\bigcap_{i \in I'} A_i$  is not smaller in the space conditional on  $\bar{A}_j$ .

Since thus  $\Pr[\bigcap_{i \in I'} A_i \mid \bar{A}_j] \geq \Pr[\bigcap_{i \in I'} A_i]$ , we see from equation (1) that  $\Pr[\bigcap_{i \in I'} A_i \mid A_j] \leq \Pr[\bigcap_{i \in I'} A_i]$ . From  $\Pr[\bigcap_{i \in I'} A_i \mid A_j] = \Pr[\bigcap_{i \in I} A_i] / \Pr[A_j]$  we derive the desired statement  $\Pr[\bigcap_{i \in I} A_i] \leq \Pr[\bigcap_{i \in I'} A_i] \Pr[A_j]$ . Applying induction to  $I'$ , we have  $\Pr[\bigcap_{i \in I} A_i] \leq \prod_{i \in I'} \Pr[A_i] \Pr[A_j] = \prod_{i \in I} \Pr[A_i]$ .

Using this negative correlation property for the set of all missing bits, we conclude that the probability  $\Pr[\bigcap_{i \in M} A_i]$  that all missing bits were flipped at least once, is  $\Pr[\bigcap_{i \in M} A_i] \leq \prod_{i \in M} \Pr[A_i] \leq (1 - n^{-4\alpha})^{n/3} \leq \exp(-n^{-4\alpha})^{n/3} = \exp(-n^{1-4\alpha}/3)$ , where we used the estimate  $(1 + x) \leq e^x$  valid for all  $x \in \mathbb{R}$ .

Consequently, with probability at least  $1 - \exp(-\Omega(n)) - \exp(-\Omega(kt)) - \exp(-\Omega(n^{1-4\alpha}))$ , there is a bit that initially has the value zero and is not flipped in the first  $t$  iterations, implying that the  $(1 + (\lambda, \lambda))$  GA needs more than  $t$  iterations to generate the optimum as mutation or crossover offspring. This high-probability statement immediately implies the claimed bound on the expected optimization time, using again  $E[T] \geq 2$  and  $E[F] = \Theta(\lambda E[T])$ .  $\square$

**Lemma 14.** *If  $\omega(1) = k \leq \sqrt{n}$ , then  $E[F] = \Omega(n \log n \min\{\frac{\exp(\Omega(r))}{\lambda r}, \frac{n^3}{\lambda}, \frac{\exp(\Omega(k))}{k}\})$ .*

*Proof.* We first analyze the progress made in an iteration starting with a search point with fitness distance between  $n^{1/8}$  and  $n^{1/4}$  and then use this information with the lower bound multiplicative drift theorem to obtain the claimed lower bound for the optimization time.

Consider an iteration starting with a search point  $x$  with  $n^{1/8} \leq d(x) \leq n^{1/4}$ . Let  $z$  be a search point among  $\{x, x', y\}$  with maximal fitness. We aim at estimating the expected progress  $E[d(x) - d(z)] = E[f(z) - f(x)]$ . Since  $\ell$  is binomially distributed, we have  $\ell < k/2$  with probability at most  $\exp(-\Omega(k))$  by the multiplicative Chernoff bound. Similarly, with probability at most  $\exp(-\Omega(k))$ , we have  $\ell > 2k$ . In this case, we have  $E[\ell \mid \ell > 2k] \leq 3k + 1$  by Lemma 1. Hence  $E[\ell \mid \ell \notin [k/2, 2k]] = O(k)$ .

Let  $\tilde{x}$  be an offspring created in the mutation phase. Let  $\tilde{g} := g(x, \tilde{x})$ . Conditioning on the outcome of  $\ell$ ,  $\tilde{g}$  has a hypergeometric distribution with parameters  $n$ ,  $\ell$ , and  $d$ . Hence  $E[\tilde{g}] = d\ell/n$ . For the mutation winner  $x'$ , note that  $g' := g(x, x') \leq \sum_{i=1}^{\lambda} g(x, x^{(i)})$ . Hence  $E[g'] \leq \lambda d\ell/n$ .

For  $\ell \notin [k/2, 2k]$ , we use the estimate that  $f(z) - f(x) \leq g'$  with probability one (note

that this estimate is fulfilled both for  $z = x'$  and  $z = y$ ). Hence we compute

$$\begin{aligned}
E[f(z) - f(x) \mid \ell \notin [k/2, 2k]] &= \sum_{i \notin [k/2, 2k]} \Pr[\ell = i \mid \ell \notin [k/2, 2k]] E[f(z) - f(x) \mid \ell = i] \\
&\leq \sum_{i \notin [k/2, 2k]} \Pr[\ell = i \mid \ell \notin [k/2, 2k]] E[g' \mid \ell = i] \\
&\leq \sum_{i \notin [k/2, 2k]} \Pr[\ell = i \mid \ell \notin [k/2, 2k]] \lambda di/n \\
&= E[\ell \mid \ell \notin [k/2, 2k]] \lambda d/n = O(k\lambda d/n).
\end{aligned}$$

Hence let us assume (and condition on) that  $k/2 \leq \ell \leq 2k$ . Then  $E[\tilde{g}] = d\ell/n \leq 2n^{-1/4}$  and thus  $\Pr[\tilde{g} \geq 20] \leq O(n^{-5})$  by Theorem 3 (a) and Theorem 4. Similarly,  $E[g'] = \lambda d\ell/n \leq 2\lambda dk/n$  and the probability that  $x'$  has a good bit at all is  $\Pr[g' \geq 1] \leq E[g'] = 2\lambda dk/n$  by Markov's inequality. If  $g' = 0$ , then  $f(x) = f(z)$ . So let us consider the case that  $g' > 0$ . Without conditioning on  $g' > 0$ , we have  $\Pr[g' \geq 20] \leq \lambda \Pr[\tilde{g} \geq 20] = O(\lambda n^{-5})$ . Hence conditional on  $g' > 0$ , this probability is at most  $O(\lambda n^{-5} / \Pr[g' \geq 1]) = O(\lambda n^{-5} / \min\{1, 2\lambda dk/n\}) = O(\max\{\lambda n^{-5}, n^{-4}/(dk)\})$ . In this rare event, we can safely estimate  $f(z) - f(x) \leq n$ , so let us turn to the more interesting case that  $1 \leq g' < 20$ . Since  $H(x, x') = \ell$ , we have  $b(x, x') \geq \ell - 19$ . Consequently,  $\ell = \Theta(k) = \omega(1)$  implies that no mutation offspring can be better than  $x$ . Let  $\tilde{y}$  be an offspring generated in the crossover phase. Let  $b_c := b(x, x', \tilde{y})$  denote the number of bad bits of  $x'$  that make it into  $\tilde{y}$ . For  $f(\tilde{y}) > f(x)$  to hold, we need that  $b_c \leq 19$ , but also that at least one good bit makes it into  $\tilde{y}$ , that is,  $g(x, x', \tilde{y}) \geq 1$ . Since  $b_c$  follows a binomial distribution with parameters  $b(x, x')$  and  $r/k$ , we have  $E[b_c] = b(x, x')r/k \geq (\ell - 19)r/k$ . Hence  $\Pr[b_c \leq 19] \leq \exp(-\Omega(r))$  by the multiplicative Chernoff bound. The expected number of good bits making it into  $\tilde{y}$  is at most  $E[g(x, x', \tilde{y})] \leq 19 \cdot (r/k)$ , hence by Markov's inequality this is also an upper bound for the probability that good bits make it into  $\tilde{y}$  at all. Putting all this together and taking a union bound over the  $\lambda$  crossover offspring, we see that (still in the case that  $1 \leq g' \leq 19$ ) the probability that some crossover offspring is better than  $x$  is at most  $\lambda \cdot (19r/k) \cdot \exp(-\Omega(r))$ ; only then we have  $f(z) > f(x)$ , however, the gain is at most 19. Consequently,  $E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge 1 \leq g' \leq 19] \leq \Pr[f(z) > f(x) \mid k/2 \leq \ell \leq 2k \wedge 1 \leq g' \leq 19] \cdot 19 \leq 19\lambda(19r/k) \exp(-\Omega(r))$ .

We thus have

$$\begin{aligned}
&E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k] \\
&= \Pr[g' \geq 1] E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge g' \geq 1] \\
&= (\lambda dk/n) (\Pr[g' \leq 19 \mid g' \geq 1] E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge 1 \leq g' \leq 19] \\
&\quad + \Pr[g' \geq 20 \mid g' \geq 1] E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge g' \geq 20]) \\
&\leq (\lambda dk/n) (1 \cdot 19^2 \lambda r \exp(-\Omega(r))/k + O(\max\{\lambda n^{-5}, n^{-4}/(dk)\})n) \\
&\leq O(\lambda^2 dr \exp(-\Omega(r))n^{-1} + \lambda^2 dkn^{-5} + \lambda n^{-4}) \\
&= O(d\lambda^2(r \exp(-\Omega(r))n^{-1} + n^{-4})).
\end{aligned}$$

Together with the exceptional case that  $\ell \notin [k/2, 2k]$ , we obtain

$$\begin{aligned}
E[f(z) - f(x)] &= \Pr[k/2 \leq \ell \leq 2k] E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k] \\
&\quad + \Pr[\ell \notin [k/2, 2k]] E[f(z) - f(x) \mid \ell \notin [k/2, 2k]] \\
&= O(d\lambda^2(r \exp(-\Omega(r)n^{-1}) + n^{-4}) + \exp(-\Omega(k))O(k\lambda d/n)) \\
&= O(\frac{d\lambda}{n}(\lambda r \exp(-\Omega(r)) + \lambda n^{-3} + k \exp(-\Omega(k)))).
\end{aligned}$$

We now use the lower bound multiplicative drift theorem (Theorem 6) to prove our claim. By Proposition 8, with high probability a run of the  $(1+(\lambda, \lambda))$  GA once encounters a search point  $x_0$  with  $d(x_0) \in [0.5n^{1/4}, n^{1/4}]$ . For this case, we give a lower bound for the expected optimization time (which implies asymptotically the same bound for the general case). Denote by  $x_t, t \geq 0$ , the sequence of search points  $x$  generated by the  $(1+(\lambda, \lambda))$  GA in the sequel (except that  $x_t$  is the optimum solution from the first point on that the optimum was found). Let  $s_{\min} := n^{1/8}$ . We just showed that  $E[d(x_{t+1}) - d(x_t) \mid d(x_t) = s] \leq s\delta$  holds for all  $s \in [s_{\min}, d(x_0)]$ , where we set  $\delta = K \frac{\lambda}{n}(\lambda r \exp(-\Omega(r)) + \lambda n^{-3} + k \exp(-\Omega(k)))$  for some absolute constant  $K$ . By Proposition 8 again, we know that

$$\Pr[d(x_t) - d(x_{t+1}) \geq 0.5s \mid d(x_t) = s] \leq \lambda(\lambda + 1) \exp(-s) \leq 0.5\delta / \ln(s). \quad (2)$$

Consequently, we may apply Theorem 6 to the random process  $(\max\{1, d(x_t)\})_{t \geq 0}$ , and learn that the expected first  $t$  such that  $d(x_t) \leq s_{\min}$  is  $\Omega(\log(n)/\delta) = \Omega(\frac{n \log n}{\lambda(\lambda r \exp(-\Omega(r)) + \lambda n^{-3} + k \exp(-\Omega(k)))})$ . Consequently,  $E[T]$  is at least this number. By (2), we also have  $E[T] \geq 2$  and thus  $E[F] = \Omega(\lambda E[T]) = \Omega(\frac{n \log n}{\lambda r \exp(-\Omega(r)) + \lambda n^{-3} + k \exp(-\Omega(k))}) = \Omega(n \log n \min\{\frac{\exp(\Omega(r))}{\lambda r}, \frac{n^3}{\lambda}, \frac{\exp(\Omega(k))}{k}\})$ .  $\square$

The following result exploits the simple fact that if in one iteration a mutation strength of  $\ell$  was used, then regardless of the population size no progress of more than  $\ell$  can be made.

**Lemma 15.** *Let  $k \leq n/4$ . Then  $E[F] = \Omega(\frac{n\lambda}{k})$ .*

*Proof.* Let  $x_0$  be the random initial search point. When  $x_t$  is defined for some  $t \geq 0$ , let  $x_{t+1}$  be the value of  $x$  after one iteration of the  $(1+(\lambda, \lambda))$  GA starting with  $x = x_t$ , unless this iteration generated the optimal solution, in this case let  $x_{t+1}$  be the optimal solution. Hence the sequence  $(x_t)_t$  describes a typical run of the  $(1+(\lambda, \lambda))$  GA until the point when an optimal solution was generated. In particular,  $T = \min\{t \geq 0 \mid d(x_t) = 0\}$ .

We use the simple argument that all offspring generated in one iteration have a Hamming distance of at most  $\ell$  from the parent. Consequently,  $E[d(x_t) - d(x_{t+1})] \leq E[\ell] = k$ , regardless of whether  $x_{t+1}$  is an optimal mutation offspring or the crossover winner. By the additive drift theorem (Theorem 5), we have  $E[T \mid x_0] \geq d(x_0)/k$ . Since the expected distance of a random search point from the optimum is  $n/2$ , the law of total expectation gives  $E[T] \geq E[d(x_0)]/k = n/2k$ . This is at least 2, so by Proposition 10, we have  $E[F] = \Omega(\frac{n\lambda}{k})$ .  $\square$

**Lemma 16.** *Let  $k \leq n/80$ ,  $\lambda \leq \exp(k/120)$ ,  $\lambda = \exp(o(n))$ , and  $\lambda = \omega(1)$ . Then  $E[F] = \Omega(\frac{n\lambda \log \log(\lambda)}{r \log \lambda})$ .*

*Proof.* We shall show that the expected fitness gain in an iteration started with a search point with fitness distance at most  $n/10$ , is  $O(r \log \lambda / \log \log \lambda)$ . Since the  $(1 + (\lambda, \lambda))$  GA by Proposition 8, here we use the assumption  $\lambda = \exp(o(n))$ , with high probability reaches once a search point  $x$  with  $f(x) \in [n/20, n/10]$ , the claim follows from the additive drift theorem (Theorem 5).

To prove the drift condition, consider one iteration of the  $(1 + (\lambda, \lambda))$  GA started with a parent individual  $x$  with  $d(x) \leq n/10$ . Let  $z$  be the value of  $x$  after one iteration, or the optimal search point if it was found as a mutation offspring (hence, as mutation winner). We show that the expected fitness gain  $f(z) - f(x)$  is at most  $O(\log \lambda / \log \log \lambda)$ . For this, we first argue that we can assume that  $k/2 \leq \ell \leq 2k$ . Indeed, we have

$$\begin{aligned} E[f(z) - f(x)] &= \Pr[\ell < k/2] E[f(z) - f(x) \mid \ell < k/2] \\ &\quad + \Pr[k/2 \leq \ell \leq 2k] E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k] \\ &\quad + \Pr[\ell > 2k] E[f(z) - f(x) \mid \ell > 2k]. \end{aligned}$$

By the multiplicative Chernoff bounds of Theorem 3, both  $\Pr[\ell < k/2]$  and  $\Pr[\ell > 2k]$  are  $\exp(-\Omega(k))$ . Since all offspring generated in one iteration (in either mutation and crossover phase) have Hamming distance at most  $\ell$  from  $x$ , we immediately have  $E[f(z) - f(x) \mid \ell < k/2] < k/2$ . By Lemma 1, we also have  $E[f(z) - f(x) \mid \ell > 2k] \leq E[\ell \mid \ell > 2k] \leq 3k + 1$ . Hence  $E[f(z) - f(x)] \leq k \exp(-\Omega(k)) + E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k] \leq O(1) + E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k]$ .

Hence we can assume for the remainder that  $k/2 \leq \ell \leq 2k$ . In this case, we argue as follows. Consider a mutation offspring  $\tilde{x}$  and let  $\tilde{g} := g(x, \tilde{x})$ . Then  $E[\tilde{g}] = \ell d(x)/n \leq \ell/10$ . The probability that  $\tilde{g} \geq \ell/5$  is at most  $\exp(-(\ell/10)/3) \leq \exp(-k/60)$  by Theorem 3 (b)<sup>2</sup> and Theorem 4. Since  $\lambda \leq \exp(k/120)$ , we see that with probability at least  $1 - \exp(-k/120)$ , all mutation offspring have at most  $\ell/5$  good bits, implying that  $g' := g(x, x')$  satisfies  $g' \leq \ell/5$ . Note that in the rare case that  $g' > \ell/5$ , which occurs with probability at most  $\exp(-k/120)$ , we still have  $f(z) - f(x) \leq g' \leq \ell \leq 2k$  with probability one, that is, this case contributes only another  $k \exp(-\Omega(k))$  to the drift.

Therefore, let us now also condition on  $g' \leq \ell/5$ . Note that this also implies that  $b' := b(x, x')$  satisfies  $b' \geq (4/5)\ell$ , since all mutation offspring have Hamming distance exactly  $\ell$  from the parent  $x$ . Consequently, all mutation offspring are worse than  $x$ , and  $z \in \{x, y\}$ .

We now analyze the result of a crossover phase. Consider a crossover offspring  $y^{(j)}$  and let  $g_j := g(x, x', y^{(j)})$ . Then  $E[g_j] \leq g'r/k \leq (\ell/5) \cdot (r/k) \leq (2/5)r$ . Let  $\Delta = \frac{2r \ln(\lambda)}{\ln \ln(\lambda)} + s$

---

<sup>2</sup>To be precise, we use here the fact that the bound of Theorem 3 (b) is also valid if both occurrences of  $E[X]$  are replaced by an upper bound for  $E[X]$ . This is a well-known fact, but seemingly a reference is not so easy to find. Hence the easiest solution is maybe to derive this fact right from Theorem 3 (b) by extending the sequence  $X_1, \dots, X_n$  of random variables by random variables that take a certain value with probability one. By this, we can artificially increase  $E[X]$  without changing the random variable  $X - E[X]$ . Hence the bound obtained from applying the Theorem to the extended sequence applies also to the original one.

for a non-negative integer  $s$ . By Theorem 3 (a),

$$\begin{aligned} \Pr \left[ \max_{j \in [1..\lambda]} g_j \geq \Delta \right] &\leq \sum_{j=1}^{\lambda} \Pr[g_j \geq \Delta] \\ &\leq \lambda \left( \frac{eE[g_j]}{\Delta} \right)^{\Delta} \leq \lambda \left( \frac{e \ln \ln(\lambda)}{5 \ln(\lambda)} \right)^{2 \frac{\ln(\lambda)}{\ln \ln(\lambda)} + s} \leq 2^{-s}. \end{aligned}$$

Consequently, by Lemma 2,

$$E \left[ \max_j g_j \right] = \sum_{t=1}^{\infty} \Pr \left[ \max_j g_j \geq t \right] \leq \frac{2r \ln \lambda}{\ln \ln \lambda} + \sum_{s=1}^{\infty} 2^{-s} \leq \frac{2r \ln \lambda}{\ln \ln \lambda} + 1.$$

Clearly, the number of surviving good bits is an upper bound on the progress  $f(z) - f(x)$ . Hence the expected progress of one iteration, conditional on the assumptions made before, is at most  $E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge |G'| \leq \ell/5] \leq \frac{2r \ln \lambda}{\ln \ln \lambda} + 1$ . Since the drift is always bounded by  $\ell \leq 2k$ , we have in fact  $E[f(z) - f(x) \mid k/2 \leq \ell \leq 2k \wedge |G'| \leq \ell/5] \leq \min\{2k, \frac{2r \ln \lambda}{\ln \ln \lambda} + 1\}$ . The unconditional drift thus is  $E[f(z) - f(x)] \leq \min\{2k, \frac{2r \ln \lambda}{\ln \ln \lambda} + 1\} + O(k) \exp(-\Omega(k)) = O(\min\{2k, \frac{2r \ln \lambda}{\ln \ln \lambda}\})$ . The additive drift theorem (Theorem 5), keeping in mind that we start with a search point with distance at least  $n/20$ , hence yields  $E[T] = \Omega(\max\{(n/20)/2k, (n/20)\frac{\ln \ln \lambda}{2r \ln \lambda}\})$ . This is at least 2, so we conclude  $E[F] = \Omega(\lambda E[T]) \geq \Omega(n \frac{\lambda \ln \ln \lambda}{2r \ln \lambda})$ .  $\square$

**Lemma 17.** *Let  $\lambda = \Theta(\lambda^*)$ ,  $k = \exp(\omega(\sqrt{\log(n) \log \log \log(n) / \log \log(n)}))$ ,  $k \leq n/2$ , and  $r = \Theta(1)$ . Then the expected runtime of the  $(1 + (\lambda, \lambda))$  GA with these parameters is  $\omega(F^*)$ .*

*Proof.* We first analyze the progress the  $(1 + (\lambda, \lambda))$  GA makes in one iteration starting with a search point  $x$  having fitness distance  $d := d(x) \in [3 \ln \ln(n)n/k, n/3] =: [d_0, d_1]$ . Let  $z$  denote the new parent individual after one iteration (which is either  $x$  or  $y$ ), or the optimal solution in case one of the mutation offspring generated in this iteration was optimal. To use a lower bound drift theorem later, we prove an upper bound for  $E[d(x) - d(z)]$ .

We first convince ourselves that it is very unlikely that a mutation offspring is better than  $x$ . This will allow us to only regard the situation that  $z \in \{x, y\}$ . For a mutation offspring  $\tilde{x}$  to be better than the parent  $x$ , more zero-bits have to flip than one-bits, that is,  $\tilde{g} := g(x, \tilde{x}) > b(x, \tilde{x}) =: \tilde{b}$ . By a simple domination argument, we see that this event is most likely for  $d(x) = n/3$ , so let us assume this for the moment. Then  $E[\tilde{g}] = k/3$  and  $E[\tilde{b}] = 2k/3$ . We have  $\Pr[\tilde{g} \geq k/2] = \exp(-\Omega(k))$  and  $\Pr[\tilde{b} \leq k/2] = \exp(-\Omega(k))$ . Consequently,  $\Pr[f(\tilde{x}) \geq f(x)] \leq \exp(-\Omega(k))$ . We thus compute

$$\begin{aligned} E[f(z) - f(x)] &\leq E \left[ \max_{\tilde{x}} \max\{0, f(\tilde{x}) - f(x)\} \right] + E[\max\{0, f(y) - f(x)\}] \\ &\leq \lambda n \exp(-\Omega(k)) + E[\max\{0, f(y) - f(x)\}] \\ &= O(n^{-2}) + E[\max\{0, f(y) - f(x)\}]. \end{aligned}$$

We proceed by analyzing the quality of the crossover winner. Let  $\tilde{x}$  be a mutation offspring and  $x'$  be the winning individual of the mutation phase. Let  $\tilde{g} = g(x, \tilde{x})$  and



$g' = g(x, x')$  be their numbers of good bits. We have  $E[\tilde{g}] = dk/n$  and  $\Pr[\tilde{g} \geq 2dk/n] \leq \exp(-(dk/n)/3) \leq 1/\ln(n)$ . Consequently,

$$\begin{aligned} E[\max\{0, \tilde{g} - 2dk/n\}] &\leq (1/\ln(n))E[\tilde{g} - 2dk/n \mid \tilde{g} \geq 2dk/n] \\ &= dk/n \ln(n) \end{aligned}$$

by Lemma 1. We have

$$\begin{aligned} g' &\leq \max_{\tilde{x}} g(x, \tilde{x}) \\ &= (2dk/n) + \max_{\tilde{x}} (g(x, \tilde{x}) - 2dk/n) \\ &\leq (2dk/n) + \sum_{\tilde{x}} \max\{0, g(x, \tilde{x}) - 2dk/n\} \end{aligned}$$

and thus  $E[g'] \leq 2dk/n + \sum_{\tilde{x}} E[\max\{0, g(x, \tilde{x}) - 2dk/n\}] = 2dk/n + dk\lambda/n \ln(n) = (2 + o(1))dk/n$ , where all summations and maxima are taken over all mutation offspring.

Consider an offspring  $\tilde{y}$  generated in the crossover phase. Let  $\tilde{g}_{\tilde{y}} := g(x, x', \tilde{y})$ . Then  $E[\tilde{g}_{\tilde{y}}] \leq E[g'] \cdot (r/k) = O(d/n)$ . Since the crossover winner  $y$  is chosen among the crossover offspring  $\tilde{y}$  such that  $f(\tilde{y})$ , and equivalently,  $d(x) - d(\tilde{y})$ , is maximal, we have  $d(x) - d(y) \leq \sum_{\tilde{y}} \max\{d(x) - d(\tilde{y}), 0\}$ , where  $\tilde{y}$  runs over all  $\lambda$  crossover offspring. Consequently,  $E[d(x) - d(y)] = O(\lambda \frac{d}{n})$  and hence also  $E[f(z) - f(x)] = O(\lambda \frac{d}{n})$ .

Building on this drift statement, we now use Witt's lower bound result for multiplicative drift (Theorem 6). Consider a run of the  $(1 + (\lambda, \lambda))$  GA. For  $t = 0, 1, \dots$ , denote by  $x_t$  the search point  $x$  at the beginning of the  $(t + 1)$ st iteration, except when the algorithm previously had generated the optimal solution, then let  $x_t$  be the optimal solution. With probability  $1 - o(1)$ , there is a  $t_0$  such that  $n/6 \leq d(x_{t_0}) \leq n/3$ . We show that in this case, we have an expected optimization time as claimed, which implies that also the unconditioned expectation is of the same order of magnitude.

For  $t = 0, 1, \dots$  define  $X_t = \max\{d(x_{t_0+t}), 1\}$ . Observe that  $X_{t+1} \leq X_t$  for all  $t \geq 0$ . Let  $s_{\min} := d_0$ . Then we have shown above that if  $X_t = s > s_{\min}$ , then  $E[X_t - X_{t+1}] \leq K\lambda s/n$  for some absolute constant  $K$ . Hence the first condition of the drift theorem is fulfilled with  $\delta = K\lambda/n$ . From Proposition 8 we know that  $\Pr[X_{t+1} \leq s/2] \leq \lambda^2 \exp(-\Omega(s)) = \exp(-\Omega(s))$ . Hence for  $n$  (and thus also  $s$ ) sufficiently large, also the second condition of the drift theorem is satisfied (with  $\beta = 1/2$ ). We may thus apply the theorem and derive that the first  $t$  such that  $X_t \leq s_{\min}$  satisfies  $E[t] = \Omega(\frac{\ln(X_0/s_{\min})}{\delta}) = \Omega(\frac{n \log(k/\log \log n)}{\lambda}) = \Omega(\frac{n \log(k)}{\lambda})$ . Note that this, naturally, is a lower bound on  $E[T]$ . Consequently,  $E[F] = \Omega(\lambda E[T]) = \Omega(n \log k) = \omega(F^*)$ .  $\square$

## 6 Conclusion

We proved that no parameter combination for the  $(1 + (\lambda, \lambda))$  GA can lead to an asymptotically better runtime on the ONEMAX test function class than the one suggested in [DDE13], where this algorithm was first proposed. We also proved that if some offspring population size  $\lambda$ , some mutation probability  $p = k/n$ , and some crossover bias  $c = r/k$  leads to the asymptotically best runtime, then  $\lambda = \Theta(\lambda^*) = \Theta(\sqrt{\log(n) \log \log(n) / \log \log \log(n)})$ ,  $k = \Omega(\lambda^*) \cap \exp(\omega(\sqrt{\log(n) \log \log \log(n) / \log \log(n)}))$ , and  $r = \Theta(1)$ .

A closer inspection of the proofs allows (in a semi-rigorous manner) to extract some hints on the parameter choice also for optimization problems beyond the ONEMAX test function class. The most clear one is that  $r = \Theta(1)$ , that is,  $pc = \Theta(1/n)$ , seems a good choice. It was argued intuitively in [DDE13] that this is a good choice, because it results in that  $\text{cross}_c(x, \text{mut}_p(x))$  has the same distribution as applying standard bit mutation to  $x$  with the standard choice of  $1/n$  for the mutation probability. This intuitive argument is somewhat imprecise due to the fact that one iteration of the  $(1 + (\lambda, \lambda))$  GA contains two selection phases, so neither the winner of the crossover phase has a standard bit mutation distribution (with rate  $p$ ), nor the winner of the crossover phase has a bits taken independently from the mutation winner with probability  $c$ . Nevertheless, as the proofs of Lemma 12 and 14 (for a large range of parameter settings) show, in many situations a super-constant value for  $r$  leads with high probability to the event that the crossover offspring takes much more “bad” bits from the mutation winner than it takes good bits. Conversely, an  $r$ -value of  $o(1)$  together with not too small value for  $k$  leads to probability of  $\Theta(r)$  for the crossover offspring being equal to the parent  $x$ , making it useless.

For the choice of  $\lambda$ , as with all population based algorithms, it is obvious that larger  $\lambda$ -value can only be beneficial if the positive effects of the large population outnumber the higher cost for a single iteration. From Lemma 14, we see that, again for broad ranges of the other parameters, we pay for a too small  $\lambda$  when making progress is difficult. A small value of  $\lambda$  decreases both the chance to find some good bits in the mutation phase and the chance that good bits are copied into a crossover offspring. This quadratic price for a small  $\lambda$  is worth the multiplicative increase of the effort of one iteration. A similar lesson could be deduced from the fitness dependent or the self-adjusting choice of  $\lambda$  in [DDE15, DD15a], which both again suggest a larger value for  $\lambda$  when being closer to the optimum, which in the ONEMAX landscape means that it is harder to find an improvement.

For the choice of the mutation probability  $p = k/n$ , the proof of Lemma 17 shows that a large  $k$  can lead to the effect that all mutation offspring look similar. In this case, the mutation phase does not gain from the large  $k$ -value, whereas in the crossover phase the crossover bias of  $c = r/k$  makes it difficult to copy good bits into the final solution.

We hope that these insights make it easier to use the  $(1 + (\lambda, \lambda))$  GA, which both in theoretical and empirical investigations showed a promising performance. We are also optimistic that the proof ideas developed in this work make future analyses of more-dimensional parameter spaces easier.

## References

- [DD15a] Benjamin Doerr and Carola Doerr. Optimal parameter choices through self-adjustment: Applying the 1/5-th rule in discrete settings. In *Genetic and Evolutionary Computation Conference (GECCO’15)*, pages 1335–1342. ACM, 2015.
- [DD15b] Benjamin Doerr and Carola Doerr. A tight runtime analysis of the  $(1+(\lambda, \lambda))$  genetic algorithm on OneMax. In *Genetic and Evolutionary Computation Conference (GECCO’15)*, pages 1423–1430. ACM, 2015.

- [DDE13] Benjamin Doerr, Carola Doerr, and Franziska Ebel. Lessons from the black-box: Fast crossover-based genetic algorithms. In *Genetic and Evolutionary Computation Conference (GECCO'13)*, pages 781–788. ACM, 2013.
- [DDE15] Benjamin Doerr, Carola Doerr, and Franziska Ebel. From black-box complexity to designing new genetic algorithms. *Theoretical Computer Science*, 567:87–104, 2015.
- [DG13] Benjamin Doerr and Leslie Ann Goldberg. Adaptive drift analysis. *Algorithmica*, 65:224–250, 2013.
- [DHK12] Benjamin Doerr, Edda Happ, and Christian Klein. Crossover can provably be useful in evolutionary computation. *Theoretical Computer Science*, 425:17–33, 2012.
- [DJW12] Benjamin Doerr, Daniel Johannsen, and Carola Winzen. Multiplicative drift analysis. *Algorithmica*, 64:673–697, 2012.
- [Doe11] Benjamin Doerr. Analyzing randomized search heuristics: Tools from probability theory. In Anne Auger and Benjamin Doerr, editors, *Theory of Randomized Search Heuristics*, pages 1–20. World Scientific Publishing, 2011. Available at [http://www.worldscientific.com/doi/suppl/10.1142/7438/suppl\\_file/7438\\_chap01.pdf](http://www.worldscientific.com/doi/suppl/10.1142/7438/suppl_file/7438_chap01.pdf).
- [FW04] Simon Fischer and Ingo Wegener. The Ising model on the ring: Mutation versus recombination. In *Genetic and Evolutionary Computation Conference (GECCO'04)*, volume 3102 of *Lecture Notes in Computer Science*, pages 1113–1124. Springer, 2004.
- [GP14] Brian W. Goldman and William F. Punch. Parameter-less population pyramid. In *Genetic and Evolutionary Computation Conference (GECCO'14)*, pages 785–792. ACM, 2014.
- [GW15] Christian Gießen and Carsten Witt. Population size vs. mutation strength for the  $(1+\lambda)$  EA on OneMax. In *Genetic and Evolutionary Computation Conference (GECCO'15)*, pages 1439–1446. ACM, 2015.
- [HY01] Jun He and Xin Yao. Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence*, 127:57–85, 2001.
- [Jan13] Thomas Jansen. *Analyzing Evolutionary Algorithms—The Computer Science Perspective*. Springer, 2013.
- [JW02] Thomas Jansen and Ingo Wegener. The analysis of evolutionary algorithms - a proof that crossover really can help. *Algorithmica*, 34:47–66, 2002.
- [LW12] Per Kristian Lehre and Carsten Witt. Black-box search by unbiased variation. *Algorithmica*, 64:623–642, 2012.

- [MB15] Vladimir Mironovich and Maxim Buzdalov. Hard test generation for maximum flow algorithms with the fast crossover-based evolutionary algorithm. In *Genetic and Evolutionary Computation Conference (GECCO'15) (Companion Material)*, pages 1229–1232. ACM, 2015.
- [OY11] Pietro Simone Oliveto and Xin Yao. Runtime analysis of evolutionary algorithms for discrete optimization. In Anne Auger and Benjamin Doerr, editors, *Theory of Randomized Search Heuristics*, pages 21–52. World Scientific Publishing, Singapore, 2011.
- [Sud05] Dirk Sudholt. Crossover is provably essential for the Ising model on trees. In *Genetic and Evolutionary Computation Conference (GECCO'05)*, pages 1161–1167. ACM Press, 2005.
- [Wit13] Carsten Witt. Tight bounds on the optimization time of a randomized search heuristic on linear functions. *Combinatorics, Probability & Computing*, 22:294–318, 2013.